DOCUMENT RESUME

ED 270 485 TM 860 368

AUTHOR Hambleton, Ronald K.; Rogers, H. Jane
TITLE Evaluation of the Plot Method for Identifying Potentially Biased Test Items.
PUB DATE 18 Feb 86
NOTE 45p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Criterion Referenced Tests; Culture Fair Tests; Difficulty Level; Estimation (Mathematics); Goodness of Fit; Grade 3; Higher Education; *Latent Trait Theory; *Mathematical Models; *Measurement Techniques; Primary Education; Research Methodology; Sample Size; Scaling; Scores; Statistical Studies; *Test Bias; *Testing Problems; *Test Items; Test Theory
IDENTIFIERS Cleveland Public Schools OH; Item Parameters; *Plot Method; University of South Florida; University of Wisconsin

ABSTRACT

This report was designed to respond to two major methodological shortcomings in the item bias literature: (1) misfitting test models; and (2) the use of significance tests. Specifically, the goals of the research were to describe a newly developed method known as the "plot method" for identifying potentially biased test items and to conduct several methodological investigations associated with applying the plot method. Following an introduction to the theory which forms the basis for the three-parameter logistic model and the model itself, the plot method first described by Shepard was presented. Advantages of the plot method were that (1) the problem of sample size was controlled for through the baseline plots; (2) the baseline plots provided a basis for interpreting the importance of particular independent variables on the invariance property of item difficulty parameter estimates; and (3) the concept of replication could replace the concept of statistical significance testing. The plot method resulted in reasonably stable determinations of potentially biased test items with small samples. Other methods with small sample sizes were not as successful. Methodological findings provided direction for future applications of the plot method. References and figures are appended. (Author/PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED270485

Evaluation of the Plot Method for Identifying Potentially Biased Test Items

Ronald K. Hambleton and H. Jane Rogers
University of Massachusetts at Amherst

Abstract

The research described in this paper was designed to respond to two major methodological shortcomings in the item bias literature: (1) misfitting test models, and (2) the use of significance tests. The goals of the present paper were (1) to describe a new method known as the "plot method" for identifying potentially biased test items, and (2) to conduct several methodological investigations associated with applying the plot method.

The plot method resulted in reasonably stable determinations of potentially biased test items with small samples. Other methods with small sample sizes were not as successful. Methodological findings provided guidance for future applications of the plot method.

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

R. K. Hambleton

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 860 368

2/18/86

EVALUATION OF THE PLOT METHOD FOR IDENTIFYING POTENTIALLY BIASED TEST ITEMS

Ronald K. Hambleton and H. Jane Rogers
University of Massachusetts, Amherst

The notion that some items in a test may be biased against specific minority groups has become a matter of substantial concern for both test users and test developers. Numerous approaches for the study of item bias, based upon standard testing technology and item response theory, have been advanced and researched in recent years (e.g., Berk, 1982; Hoover & Kilen, 1984; Ironson & Subkoviak, 1979; Shepard, 1981; Shepard, Camilli, & Averill, 1981; Shepard, Camilli, & Williams, 1984).

The research described in this study was prompted by two problems with current research on item bias. First, while the use of item response models appears to hold considerable potential for detecting biased items (Ironson, 1982; Subkoviak, et al., 1984), many researchers are using the less defensible one-parameter item response model. Perhaps this particular choice of model is made because of the wide availability of a straightforward computer program (BICAL) to carry out one-parameter model analyses. The typical strategy is to compare item difficulty values obtained in two samples (for example, Blacks and Whites) and label items as biased (or, in need of careful scrutiny) if statistically significant differences are obtained. But, the one-parameter model is highly restrictive and so model-data fit may be

---

[1] The authors are grateful to the Cleveland Public Schools, and to Gail Ironson, University of South Florida, and Michael Subkoviak, University of Wisconsin, for providing the test data that were used in this study.

poor (Hambleton & Murray, 1983). Seldom do researchers check model-data fit in each group of interest prior to conducting bias studies. When the fit is poor, comparisons of item difficulty values are not very useful because the item difficulty values provide inaccurate information about the functioning of test items in the minority and majority groups. Poor model fit may result in the labeling of misfitting items as biased when the misfit is actually due to the failure of the one-parameter model to account for variation in item discriminating power or the guessing behavior of low-ability examinees.

A second problem arises when significance tests are used in item bias studies. Unfortunately, the results of significance tests are influenced by examinee sample size. It is well known that there is a positive relationship between the number of items detected as biased and examinee sample size. With an examinee sample size of 100, perhaps no test items will be identified; with a sample size of 5000, it is possible that nearly all test items will be identified as "biased" (see for example, Hambleton & Murray, 1983). A second problem is that the sampling distributions of many of the bias statistics are unknown so that proper hypothesis testing cannot be carried out. Alternatives are needed for studying bias that do not use significance tests, and/or are not based upon faulty distributional assumptions.

The research described in this paper was designed to address the problems introduced above. Specifically, the goals of the research were (1) to describe a new method known as the "plot method," for identifying potentially biased test items, and (2) to conduct several

methodological investigations associated with applying the plot method. The new method uses the three-parameter logistic model (Hambleton & Swaminathan, 1985) though other item response models might be used too. The plot method can overcome the two problems mentioned earlier as well as offer several advantages, and therefore the research seemed worthwhile. Many readers may not be familiar with the three-parameter logistic model and so an introduction to the theory from which the three-parameter model was derived and the model itself will be provided next.

Item Response Theory

The three-parameter logistic model is one of the class of mathematical models called item response models. These models derive from a relatively new approach to test theory, know as item response theory. Item response theory has several important advantages over classical test theory which have provided strong impetus for its development and implementation. In the last decade, it has become the dominant topic of research in measurement, and applications continue to increase in a wide variety of educational, industrial, and professional settings (Hambleton & Swaminathan, 1985; Lord, 1980).

Item response theory is based on the assumption that observed performance on the items in a test can be substantially accounted for by a set of underlying, unobservable factors called traits or abilities. In most applications at present, a single underlying factor is assumed. It is further assumed that for each item of the test, the probability

of a correct answer is a monotonically increasing function of this ability. That is, as ability increases, so does the probability of success. The mathematical function specifying this relationship is called the item characteristic function and the curve it yields, the item characteristic curve.

Item characteristic curves vary according to the characteristics of the item. Items may be described (for example) by one, two, or three parameters, representing item difficulty, discrimination, and guessing. The number of parameters incorporated in the model identifies it as a one-parameter (difficulty parameter only), a two-parameter (difficulty and discrimination parameters), or a three-parameter model (difficulty, discrimination, and guessing parameters). The one-parameter model assumes that items vary only in difficulty; all items are equally discriminating and no guessing occurs. The two-parameter model, while allowing items to vary in difficulty and discrimination, assumes that there is no guessing. The three-parameter model allows items to vary in difficulty, discrimination, and guessing.

Each item response model specifies a particular item characteristic function, depending on the number of item parameters incorporated. The three-parameter model has the most general mathematical form of the models in common use; it reduces to the two- or one-parameter model when further restrictive assumptions are made. The item characteristic function for the three-parameter model is given by the equation

$$P_g(\Theta) = c_g + (1-c_g)\frac{e^{Da_g(\Theta - b_g)}}{1+e^{Da_g(\Theta - b_g)}}, \quad g=1, 2, \ldots, n$$

where:

$P_g(\theta)$ = the probability that a randomly-chosen examinee with ability level θ answers item g correctly,

b_g = the item g difficulty parameter,

a_g = the item g discrimination parameter,

c_g = the lower asymptote of an item characteristic curve representing the probability of success on item g for low-ability examinees,

D = 1.7 (a scaling factor), and

n = the number of items in the test.

Estimation of item and ability parameters in the three-parameter logistic model is most often performed using maximum likelihood methods. That is, estimates which maximize the probability of obtaining the observed results are sought. LOGIST (Wingersky, Barton, & Lord, 1982) is the most commonly used computer program for three-parameter model estimation. However, the maximum likelihood procedures used in LOGIST do not guarantee that parameter estimates will remain within acceptable bounds. Bayesian procedures appear to offer a promising alternative (Hambleton & Swaminathan, 1985), although there is no computer program presently available which incorporates these methods.

The measurement scales for reporting item and ability parameter estimates are quite arbitrary. It is probably most common to scale either the item difficulty parameter estimates or the ability estimates to a mean of zero and a standard deviation of one (though other means and standard deviations are sometimes chosen to avoid negatives and decimals). Once this scaling is completed, the other model parameter estimates are rescaled too so that the probabilities associated with

examinees giving correct answers to test items on the original scale remain the same on the new transformed scale (see Hambleton & Swaminathan, 1985, for details).

Item response theory is based on strong assumptions which must be shown to hold, at least to an adequate degree, before the advantages of the IRT model of interest accrue. Assessing model-data fit is an essential step in any application of item response theory. Many techniques have been proposed for the investigation of goodness-of-fit; at present, however, the user must resort to evidence from a variety of less than perfect sources, since there seem to be no entirely adequate tests of model fit. Yet, given the acceptability of the unidimensionality assumption and the appropriateness of the item characteristic function, item response theory yields many important results unobtainable under classical test theory.

The main advantage is the invariance of parameter estimates. Under classical test theory, estimates of item difficulty and discrimination are dependent on the particular sample of individuals taking the test, and ability estimates are dependent on the particular set of items administered. Generalization of results to groups of people or items which differ to any substantial degree from the original groups is not possible. Item response theory overcomes this problem by providing ability estimates which incorporate information about the items and item parameters which incorporate information about the sample of examinees. By taking into account this information, parameter estimates which are independent of the testing situation are obtained.

A further problem of classical test theory is that it assumes that the standard error of measurement is the same for all examinees. However, it is often true that high and low ability examinees will be measured less precisely than middle ability examinees. Item response theory provides an estimate of standard error -- that is, precision of estimation -- for each examinee. In addition, once parameter estimates are known, item response theory allows the prediction of future performance. This knowledge is useful in adapting tests to the ability level of the examinee. Classical test theory cannot provide such information.

In summary, IRT models are especially attractive for item bias research because the invariance property of item parameter estimates means that possible ability differences between the groups of interest (e.g., Blacks and Whites) will not serve as a confounding variable when interpreting item bias results. The confounding of group ability differences and item bias results is a common problem with classical item bias studies. Essentially, IRT models make it possible to compare the performance levels of the two or more groups of interest (using the item characteristic curve estimated for each group) at points along the ability scale continuum. When the differences in the ICCs are small, no item bias is present. When the differences in the item characteristic curves are greater, explanations for the differences are sought. One explanation is that the test item is "biased" against the lower-performing group. Possibly the test item includes some unfamiliar language for this group or describes an unfamiliar situation. In any case, item characteristic curves (ICCs) provide an

excellent basis for conducting item bias studies if, of course, the ICCs fit the test data under investigation. To enhance model-test data fit and thereby reduce the problem in our work, the three-parameter logistic model was selected for subsequent use in our item-bias research.

Description of the Plot Method

The precise origin of the plot method is unknown to us, though the general approach using non-IRT concepts was described by Angoff (1982). Hambleton (1982), Hambleton and Murray (1983), and Hambleton, Martois, and Williams (1983) described the plot method in their research papers on goodness-of-fit measures for IRT models. However, it seems likely that Shepard (1981) was the first researcher to describe the general method referred to in this paper as the "plot method."

Basically, the following steps are followed in applying the method:

1. Choose the independent variable of interest for the item bias study (e.g., sex, race, geographic region, etc.). Form two groups (e.g., Males and Females) and label them "A" and "B".
2. Count the number of individuals in each group; draw a random sample from the larger group so that both groups (A and B) are of the same size.
3. Split both groups in half to form four equal-sized subgroups (A1, A2, B1, B2).
4. Conduct a three-parameter model analysis on each of the four subgroups (Wingersky et al., 1982) to obtain item and ability parameter estimates.

5. Scale the b-values in each analysis to a mean of zero and a standard deviation of one (or any common mean and standard deviation).

6. Plot the b-values from A1 and A2, and B1 and B2, to provide baseline information on the amount of scatter to be expected in the parameter estimates due to factors such as sample size and model-data misfit. A1 and A2, and B1 and B2, are randomly equivalent samples. b-values are plotted because they are more stable than the a-values (though the a-values could be plotted too).

7. Plot the b-values from A1 and B1, and A2 and B2, to determine if the amount of spread in the plots differs from the baseline plots obtained at step 6. If they do differ, then the independent variable (or a variable confounded with it) is influencing the b-values. A comparison of the A1 and B1, and A2 and B2 plots, permits the researcher to check the replicability of the findings.

8. Plot the differences A1-A2 (the differences in item difficulty estimates in the two samples, A1 and A2) and B1-B2 (the differences in item difficulty estimates in the two samples, B1 and B2) and compare to the plot of A1-B1 and A2-B2. If the plots differ, identify the test items showing consistently large differences in the A and B samples. These items are the ones that may be biased against one of the groups.

The main advantages of the plot method appear to be that (1) the problem of sample size is controlled for through the baseline plots (with small sample sizes the plots are more circular in shape), (2) the baseline plots provide a basis for interpreting the importance of particular independent variables on the invariance property of item difficulty parameter estimates, and (3) the concept of replication can replace the concept of statistical significance testing.

There are several variations on the above method. For example, items which on a priori grounds appear to be "biased" can be removed prior to step 4. With ability estimates in hand that are not influenced by potentially biased items, the potentially biased items can be returned to the analysis, and treating the ability estimates as

known (fixed), the complete set of item parameter estimates can then be obtained. In fact, this variation at step 4 was applied to the data described below. The variation seems especially useful when the ratio of the number of potentially flawed test items to total test length is high. In this case, the potentially biased test items can "contaminate" the ability estimates and make the overall bias analysis less sensitive.

Insert Figures 1, 2, and 3 about here.

Figures 1, 2, and 3 contain the results of our race bias study on a 50-item college vocabulary test administered to 2030 students (Whites = 1022, Blacks = 1008). These data were kindly provided by Michael Subkoviak and Gail Ironson.[2] These data are especially interesting because 10 of the 50 vocabulary items were included in the test because they would be "biased" against White students. These Black vocabulary words were: (with the meaning in brackets) fro (bush), member (black person), butch (lesbian), crib (apartment), kicks (shoes), clean (chic), boot (blood), greasing (eating), hog (car), and player (pimp).

Figures 1a and 1b provide an indication of the expected variability in item difficulty estimates across randomly-parallel samples of the same race and size. The scatter of points is influenced by the sample size and the model-test data fit. If, in this study,

---

[2] We are indebted to the Computing Center at the University of Wisconsin for reprocessing and rescoring the test data and doing the work at no charge.

race is not a factor in test performance, the plot of item difficulty estimates in the Black and White samples should look similar to the plots of the item difficulty estimates in the two Black samples (Figure 1a) or in the two White samples (Figure 1b). Figure 2a shows a subset of test items to be operating very differently in the Black and White samples. The same pattern is observed in the second independent Black and White samples (see Figure 2b). Clearly, race (or a variable confounded with race) is influencing the item parameter estimates, and hence there is the strong possibility of biased items in the test. In actual fact, all 10 items were grossly biased against the Whites (the word "butch" was the least biased of the vocabulary items, but even here there was an average difference of 1.17 standard deviations in the Black and White b-values), and it was extremely easy to detect them. An additional two items also showed a consistent tendency to be answered differently by Blacks and Whites.

Figure 3b shows that these Black-White item difficulty differences were also very consistent. The test items showing consistently large differences between the Black and White samples are the ones which are labelled "potentially biased" and are investigated further. These test items appear in the top right corner and the bottom left corner of the plot in Figure 3b. After identifying potentially biased items, however, it remains important to review the full set of item statistics (b, a, c, and associated errors) to determine the item statistics most responsible for the item characteristic curve differences and to investigate poor estimation as the source of the problem. Figure 3a shows that the differences B1-B2 and W1-W2 are uncorrelated, which is

the result that should be obtained when parameter estimation is being done correctly. Essentially, the variables being correlated are errors associated with parameter estimation.

Our original plan was to work with the Subkoviak-Ironson data set extensively in our research, but the ten biased items were so easy to detect that we preferred to carry on our research with a different set of data.

Methodological Investigations of the Plot Method

Introduction

In this section of the paper several methodological investigations of the plot-method will be described. The test data used were provided by the Cleveland Public Schools. The actual test studied was a 46-item criterion-referenced test that was administered to (approximately) 1200 grade 3 children in the Spring of 1983. The test seemed especially interesting in our item bias research because it was not in final form. A number of potentially biased items might be expected. Usually item bias studies are conducted on tests which have already been carefully screened (e.g., standardized achievement and aptitude tests) and so the task of finding biased test items is considerably more difficult.

There were only 296 Whites in the total sample and so subgroups of Blacks and Whites were formed that consisted of 148 individuals. Blacks were randomly drawn from the larger pool of available Black students so that the sizes of the Black subgroups and the White subgroups were equal.

Insert Figures 4, 5, and 6 about here.

Because of the modest sample sizes and because the test items were in draft form, our decision was to combine the total examinee sample (Blacks and Whites) first and estimate ability scores using 29 test items judged (a priori) to be unbiased. In this way, ability scores using non-biased test items, and stable item parameter estimates obtained from nearly 1200 examinees, could be obtained. These ability scores were then treated as known (fixed) and used to calibrate the item parameter estimates for the 46 test items within each subgroup of examinees. Such a modification to the application of the plot method can contribute substantially to the precision of the item parameter estimates in the rather small sized subgroups.

The plots in Figures 4, 5, and 6 provide the basic information for the bias study. They indicate that there are only four potentially biased items. Figures 5a and 5b are quite similar and not very different from Figures 4a and 4b. Therefore, race does not seem to be a major factor in item performance. Figure 6b shows that only four test items showed a consistent b-value difference of over .75 standard deviations in the Black and White samples.

Choice of Scaling Method

The first investigation concerned the choice of scaling method to be applied to the b-values prior to preparing the plots. Since the b-value scale is arbitrary (up to a linear transformation), it is

necessary to place the b-values from separate analyses (e.g., Blacks and Whites) on a common scale before any comparisons can be made. On the surface, the task seems easy: Each set of b-values can be transformed to a new scale with mean = 0 and standard deviation = 1 (or any other common mean and standard deviation).

A problem is that some b-values may be very large (e.g., in the analysis described in the last section, several b-values exceeded 60!) and therefore they exert a tremendous influence on the mean and standard deviation. With a very large standard deviation, most of the scaled b-values become very homogeneous and subsequent analyses become difficult to carry out. This problem perhaps can be solved by choosing some arbitrarily large value for the very big b-values (we chose $\pm$ 3.5) and/or removing these items from the calculations of means and standard deviations.

A second problem arises because some of the large b-values may be very poorly estimated. It is undesirable for these poorly estimated b-values (large or small) to have as much influence on the scaling as b-values which are more precisely estimated. Stocking and Lord (1983) provide a new method for scaling two sets of b-values to a common scale by considering the standard errors associated with the b-values. The effect of the new scaling method was studied in this part of our work.

Insert Figures 7 and 8 about here.

Figures 7 and 8 show the results of applying the Stocking-Lord scaling method. The similarity of the plots in Figures 7a, 7b, 8a, and

8b, suggest that race is not an important factor in item performance. For this one set of test data at least, the special effort to place the results on a common scale using the complicated Stocking-Lord procedures made no difference to our interpretations. Of course, their method may be more useful with other types of analyses.

Choice of Cut-off Points

Insert Figures 9 and 10 about here.

While studying plots can be helpful in identifying subsets of test items which show consistent differences in statistical properties between two groups, guidelines for interpreting these differences are not available at the present time. We would certainly be unwilling to do significance testing on these differences because, among other things, questionable assumptions about distributions would need to be made. Figures 9 and 10 show the differences in b-values for the white samples (a) and black samples (b). These sampling distributions which reflect chance differences principally due to the choice of sample size can be used to set cut-off points which can be applied to the Black and White b-value differences shown in Figures 9c, 9d, 10c, and 10 d. For example, a researcher could select points on the scale beyond which only about 5% of the b-value differences fall. These critical points could then be used to interpret the Black-White b-value differences. When race is a factor in item performance, many test items might be expected to have b-value differences which exceed the critical values.

Figure 10 differs from Figure 9 in that standardized b-value differences are plotted instead. The possible advantage of plots of standardized b-value differences is that the errors associated with the b-values under study can be considered too. Large b-value differences then are taken less seriously when the errors associated with the corresponding b-values are large. The formula for a standardized b-value difference is:

$$\frac{b_{Bi} - b_{Wi}}{\sqrt{SE^2(b_{Bi}) + SE^2(b_{Wi})}} \quad \text{where } i = 1, 2$$

and b_{Bi} and b_{Wi} are the b-values in sample i (i = 1, 2) for the Black and White subgroups, respectively, and $SE^2(b_{Bi})$ and $SE^2(b_{Wi})$ are the corresponding (squared) errors associated with the b-value estimates. Again, these data in Figure 10 reveal almost no biased items in the test. All of the distributions have about the same characteristics. If race were a factor, the distributions in (c) and (d) would differ somewhat from the distributions in (a) and (b) in Figures 9 and 10.

Comparison of Item Bias Methods

One criticism that can be made of the plot-method is that only one item statistic (item difficulty) is used in identifying potentially biased test items. Item difficulty might not vary in two groups of interest but the test items could vary in (say) discriminating power. On the other hand, the b-values are (usually) estimated with more precision than the a-values (Lord, 1980) and therefore (possibly) their

use in item bias studies could lead to more stable information about item bias.

In this phase of the work, the stability of two popular IRT item bias methods (which use all three item statistics in the three-parameter model) with the plot method was compared. The area method (Rudner, et al., 1980) and the squared difference method (Linn et al., 1981) were considered in the research. In the "area method," the area between the item characteristic curves in the two groups of interest over an interval of interest on the ability scale is used as an estimate of item bias. Of course, the minimum bias is achieved when the two curves are equivalent. Then, the item bias is zero. The more different the curves, the larger the area, and the more biased the item is assumed to be. The " squared difference method" is defined over the same interval on the ability scale; however, the square root of the sum of the squared differences between the two item characteristic curves at fixed intervals (usually .01) is used as the measure of item bias.

The second author prepared a computer program to compute the item bias statistics using the area and squared difference methods. Calculations were carried out on the ability scale between -3 and 3. Prior to computing the item bias statistics, corresponding sets of item statistics were placed on a common scale, using a method described by Linn, et al (1981). Of interest in the study was the level of replicability of the item bias statistics for the three methods across independent samples.

The same subgroups identified earlier (B1, B2, W1, W2) were used in this analysis. Again, though the sample sizes for the item

The results in Table 1 show that the plot method produced considerably more stable results. In fact, the results for the other two methods are quite unacceptable. Their stability is nearly zero.

Insert Table 1 about here.

However, no attempt should be made to generalize the findings since they almost certainly are limited to small-sample item bias studies.

Conclusion

It is difficult at this time to draw any definitive conclusions about the plot method for identifying potentially biased items. Certainly the method has some intuitive appeal and there is evidence that it can be successfully applied.

Though the Subkoviak-Ironson data set provided a very simple check on the method, it can be stated that all ten flawed items were easily identified (along with two additional items). Of course the level of bias in the items in the vocabulary test probably far exceeded the level of bias that could be expected in many achievement and aptitude tests, even tests that were being field-tested.

The methodological investigations produced a number of useful results. Apparently the Stocking-Lord method for placing item parameter estimates on a common scale did not produce results that differed from the use of standard score equating (where the very large values are eliminated from the calculations) for the type of

application of b-values described in this paper. This result was obtained even though the standard errors associated with the b-value estimates ranged widely and in some cases the standard errors were very large. If the Stocking-Lord scaling method was going to produce different results from standard score equating, it would have been under these conditions.

The sampling distribution of b-value differences under the null-hypothesis that the two groups are identical (B1, B2, W1, W2), provides a helpful way for obtaining cut-off points for identifying potentially biased test items since no distributional assumption must be made. The actual distribution can be produced though the shortness of a test may limit the usefulness of this distribution somewhat. The cut-off points can be positioned at the $P_{.025}$ and $P_{.975}$ points of the distribution, or alternatively, at breaks in the distribution of differences. This choice of cut-off score will result in a conservative approach to identifying biased test items, however, since often the amount of true bias in test items may be small. Still, the sampling distribution does provide a useful frame of reference for determining the size of a significant b-value difference.

Finally, evidence for the stability of the item bias results obtained from the plot method is encouraging when compared to the two other IRT item bias methods. Of course, it is unknown how these other methods would fare with larger samples. Our prediction is that the differences in the methods would be far less dramatic. In any case, the results reported in this paper seem to suggest that with small

samples, perhaps methods which focus on the most stable of the item parameter statistics should be given preference.

In summary, additional investigations of the plot method would seem to be warranted based upon the results reported in this paper. In addition, some of the methodological research reported here provides direction for applying the plot method and interpreting the results. Our next step will be to validate the method by addressing the agreement between test items identified as potentially biased by the plot method and test items identified as potentially biased via the use of judgmental review methods (Berk, 1982). Agreement in the results from these two very different methods would lend credibility to both approaches for identifying potentially biased test items.

References

Angoff, W.H. (1982). Use of difficulty and discrimination indices for detecting item bias. In R.A. Berk (Ed.), Handbook of methods for detecting test bias. Baltimore, MD: the Johns Hopkins University Press.

Berk, R.A. (Ed.) (1982). Handbook of methods for detecting test bias. Baltimore, MD: The Johns Hopkins University Press.

Hambleton, R.K. (1982). Applications of item response models to NAEP mathematics exercise results. Final Report -- ECS Contract No. 02-81-20319. Denver, CO: Educational Commission of the States.

Hambleton, R.K., Martois, J.S., & Williams, C. (1983). Detection of biased test items with item response models. Paper presented at the annual meeting of AERA, Montreal.

Hambleton, R.K., & Murray, L.N. (1983). Some goodness of fit investigations for item response models. In R.K. Hambleton (Ed.) Applications of Item Response Theory. Vancouver, B.C.: Educational Research Institute of British Columbia.

Hambleton, R.K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Hingham, MA: Kluwer-Nijhoff.

Hoover, H.D., & Kilen, M.J. (1984). The reliability of six item bias indices. Applied Psychological Measurement, 8, 173-181.

Ironson, G.H. (1982). Use of chi-square and latent trait approaches for detecting item bias. In R. Berk (Ed.), Handbook of Methods for Detecting Test Bias. Baltimore, MD: The Johns Hopkins University Press.

Ironson, G.H., & Subkoviak, M. (1979). A comparison of several methods of assessing item bias. Journal of Educational Measurement, 16, 209-225.

Linn, R.L., Levine, M.V., Hastings, C.N., & Wardrop, J.L. (1981). Item bias in a test of comprehension. Applied Psychological Measurement, 5, 159-173.

Lord, F.M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates.

Rudner, L.M., Getson, P.P., & Knight, D.L. (1980). A Monte Carlo comparison of seven biased item detection techniques. Journal of Educational Measurement, 17, 1-10.

Shepard, L.A. (1981). Identifying bias in test items. In B.F. Green (Ed.), Issues in Testing: Coaching, Disclosure and Ethnic Bias. San Francisco: Jossey-Bass.

Shepard, L., Camilli, G., & Averill, M. (1981). Comparison of procedures for detecting test-item bias with both internal and external ability criteria. Journal of Educational Statistics, 6, 317-375.

Shepard, L., Camilli, G., & Williams, D.M. (1984). Accounting for statistical artifacts in item bias research. Journal of Educational Statistics, 9, 93-138.

Stocking, J.L., & Lord, F.M. (1983). Developing a common metric in item response theory. Applied Psychological Measurement, 7, 201-210.

Subkoviak, M.J., Mack, J.S., Ironson, G.I., & Craig, R.D. (1984). Empirical comparison of selected item bias detection procedures with bias manipulation. Journal of Educational Measurement, 21, 49-58.

Wingersky, M.S., Barton, M.A., & Lord, F.M. (1982). LOGIST user's guide. Princeton, NJ: Educational Testing Service.

Table 1

Stability of Item Bias Statistics
Across Independent Samples
(Black vs. White, N = 46 items)

Statistic	B-W (Group 1) vs. B-W (Group 2) Correlation	Consistency Overall	P1	P2
Plot Method	.643	.935	.50(06)	.50(06)
Area Method	.159	.826	.33(06)	.33(06)
Squared Difference Method	.028	.783	.00(06)	.00(04)

Overall = proportion of test items identified consistently as "biased" or "unbiased" in both samples.

P1 = proportion of test items identified as "biased" in sample 1 which were also identified as "biased" in sample 2. P1 reflects the consistency of biased items from sample 1.

P2 = proportion of test items identified as "biased" in sample 2 which were also identified as "biased" in sample 1. P2 reflects the consistency of biased items from sample 2.

The numbers in brackets correspond to the numbers of test items exceeding the cut-off score for a particular group (under the P1 column the group is sample 1; under the P2 column the group is sample 2) and a particular method.

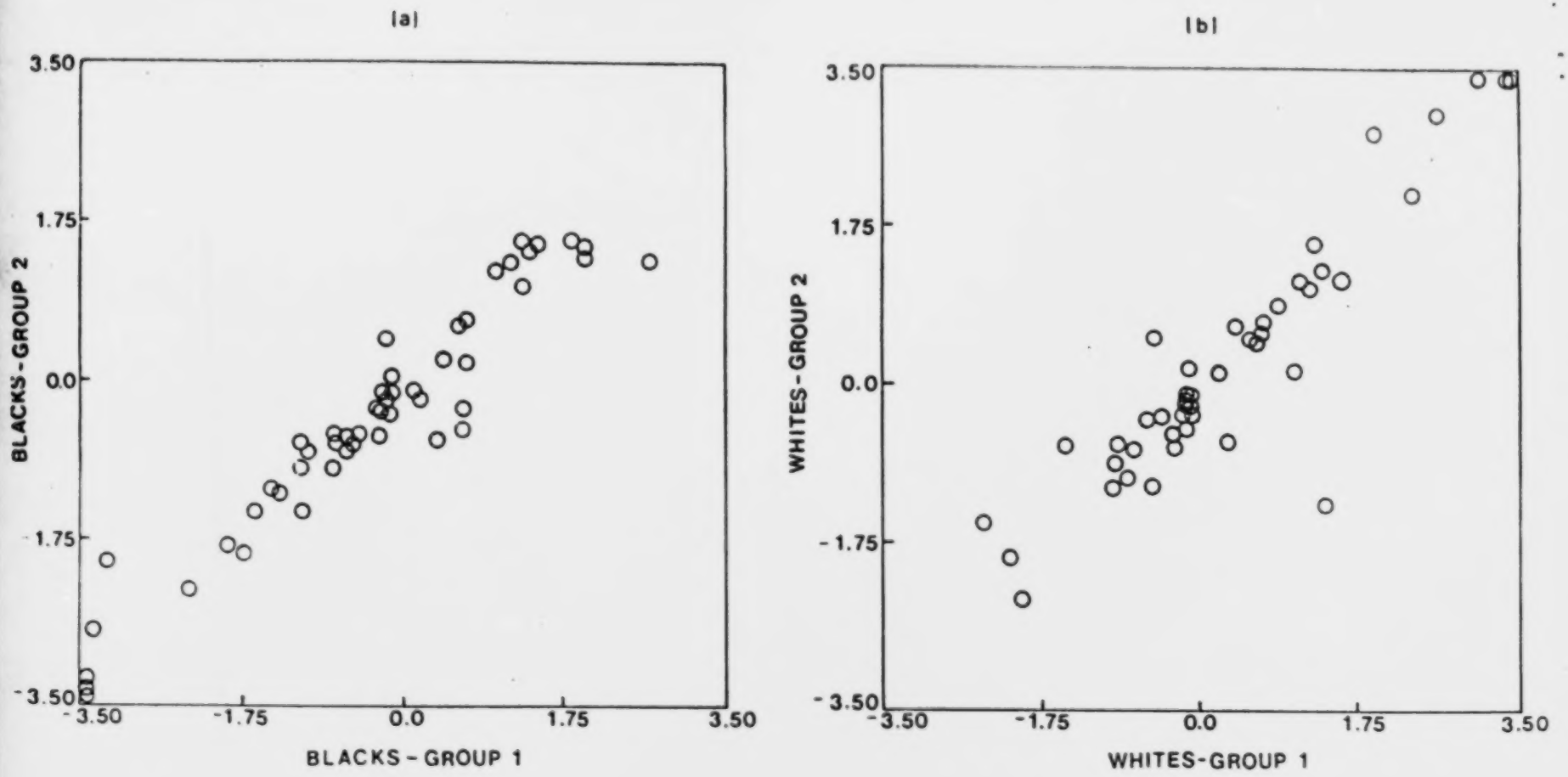

Figure 1. Plot of item difficulty estimates in two equivalent black samples (N= 504) in (a) and white samples (N = 511)in (b).

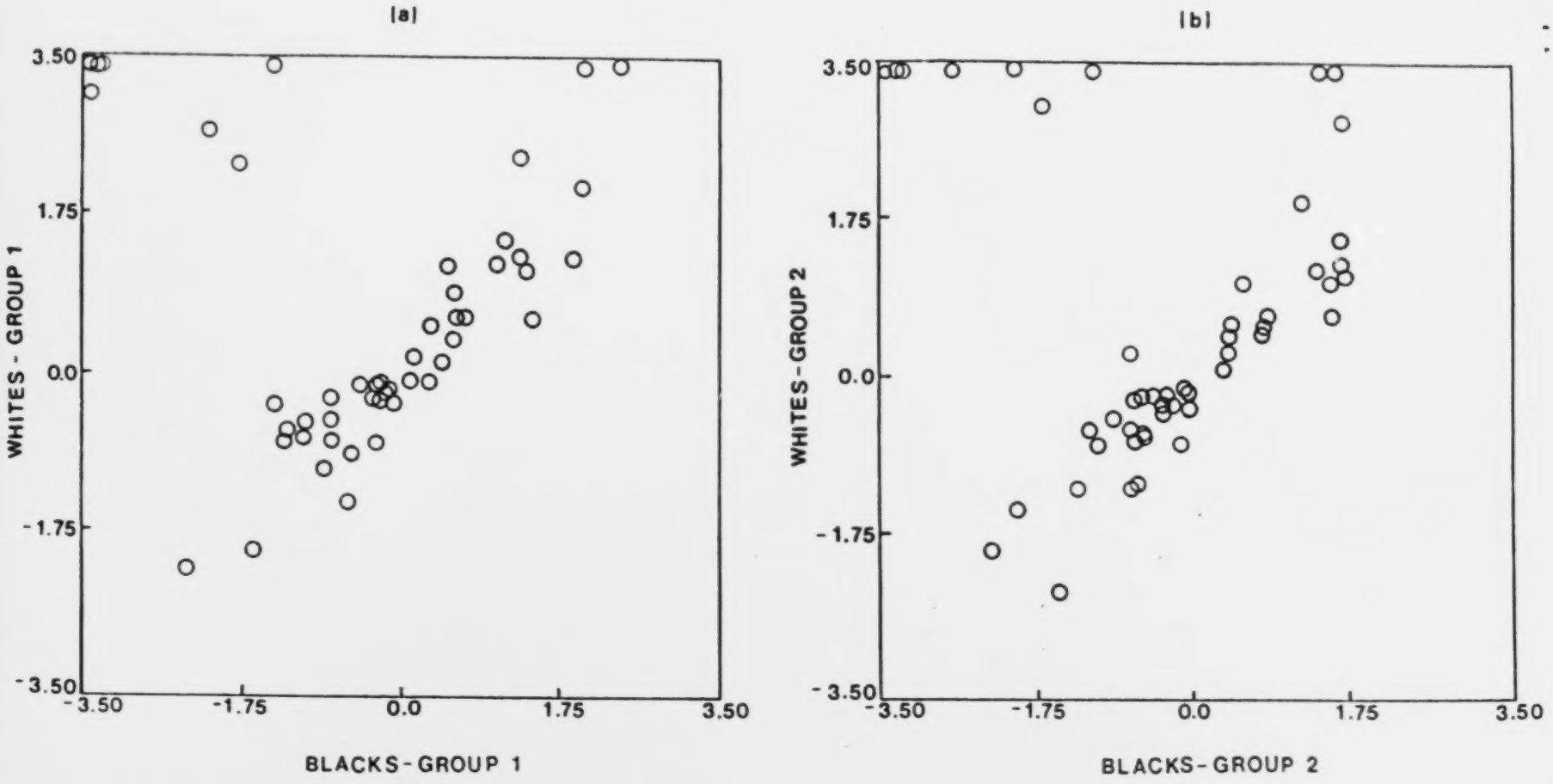

Figure 2. Plot of item difficulty estimates in the black and white samples (Sample 1 in a, Sample 2 in b).

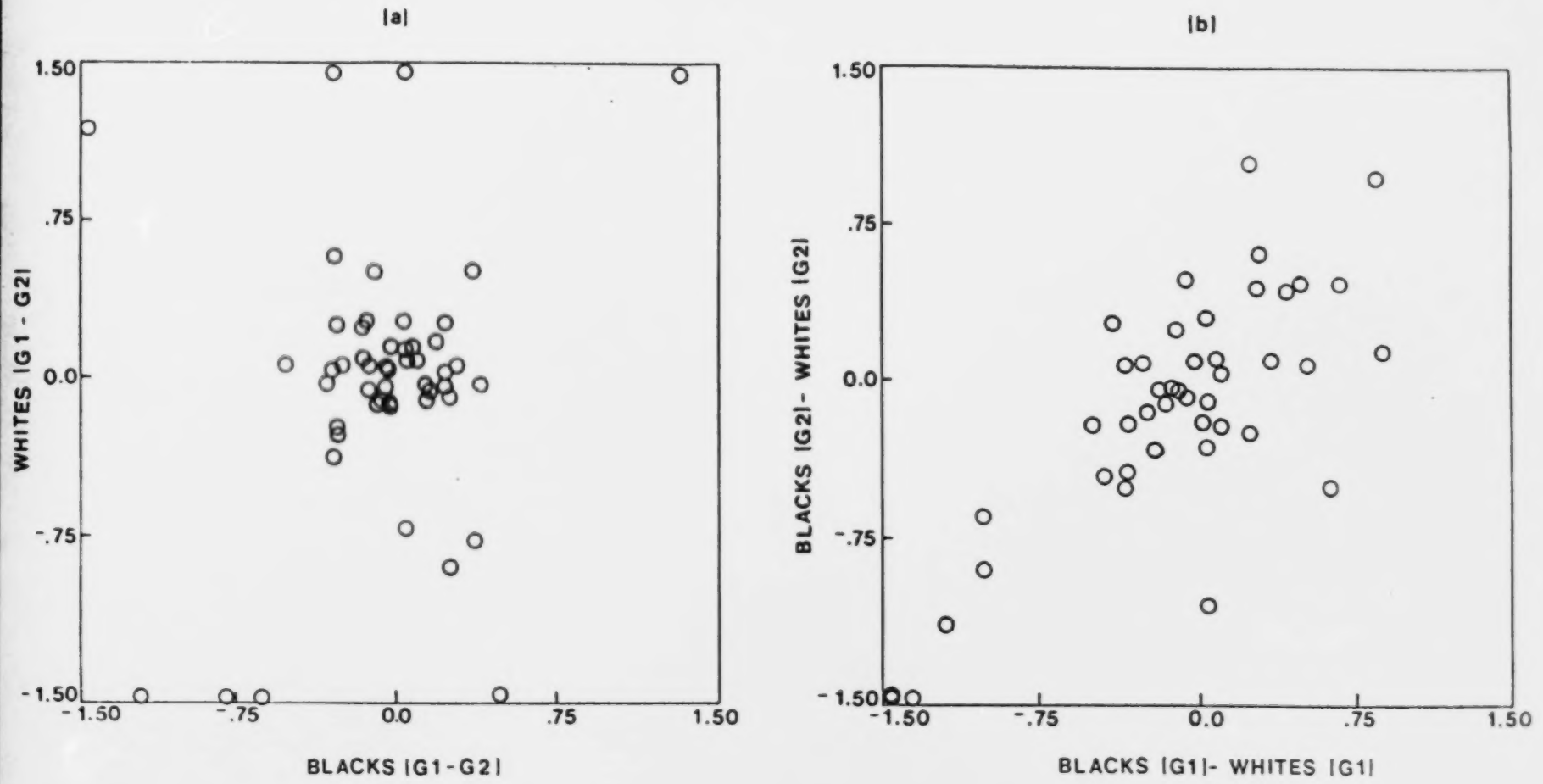

Figure 3. Plot of differences in item difficulty estimates (B1 - B2 vs. W1 - W2 in a, and B1 - W1 vs. B2 - W2 in b).

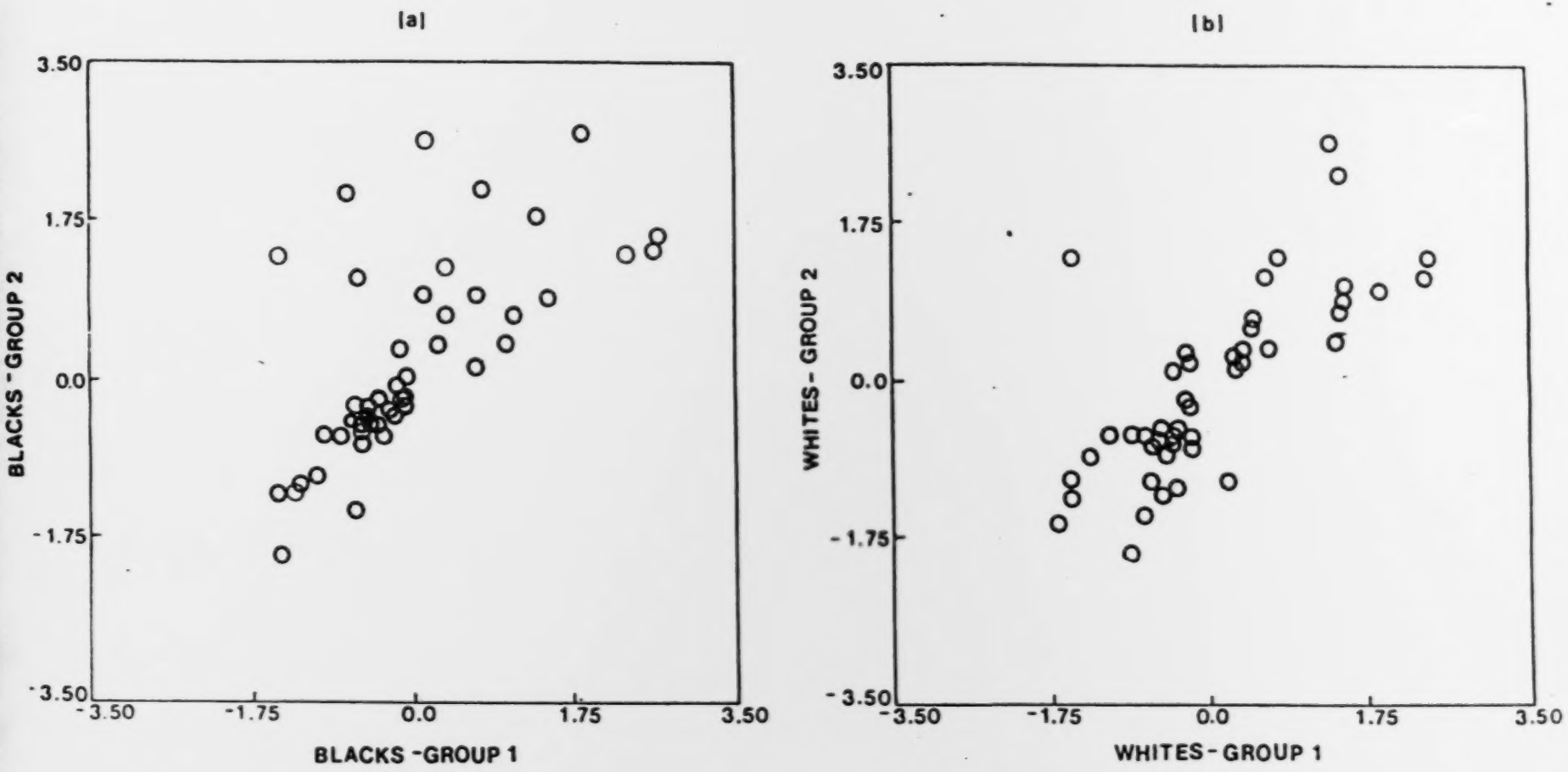

Figure 4. Cleveland data plots of item difficulty estimates in two equivalent black samples (N=148) in (a) and white samples (N=148) in (b).

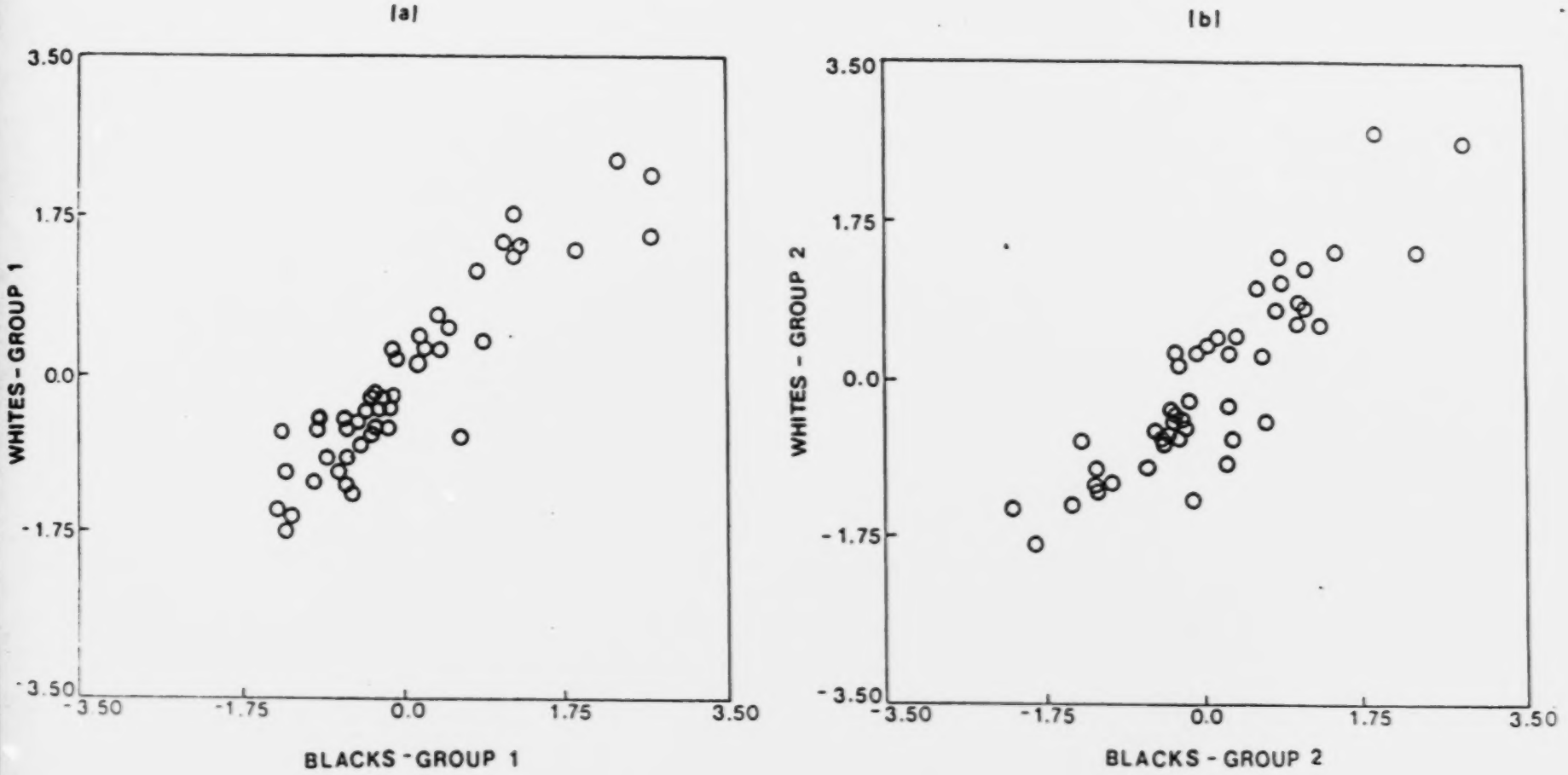

Figure 5. Cleveland data plots of item difficulty estimates in the black and white samples (Sample 1 in a, Sample 2 in b).

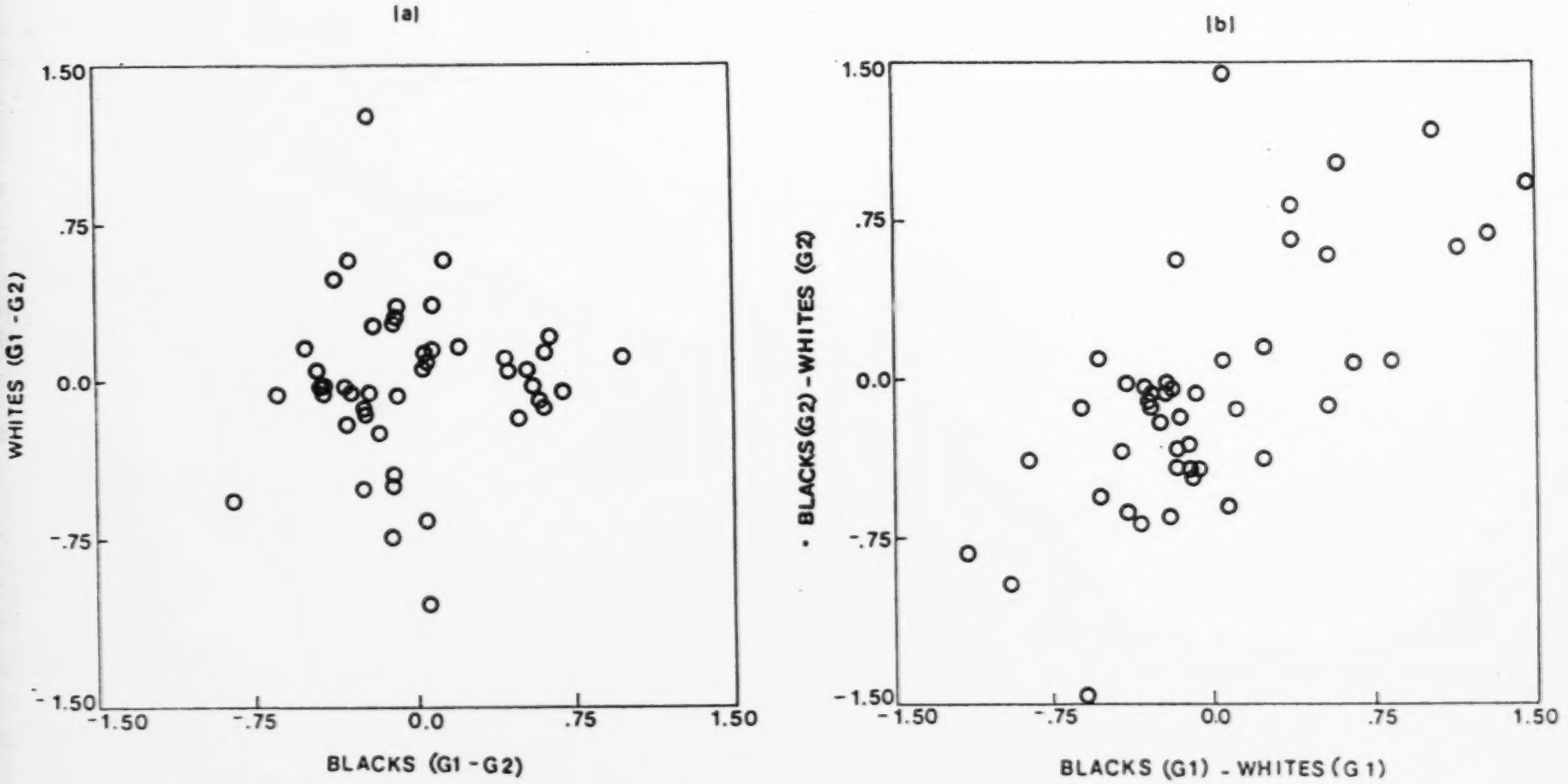

Figure 6. Cleveland data plots of differences in item difficulty estimates (B1-B2 vs. W1-W2 in a, B1-W1 vs. B2-W2 in b).

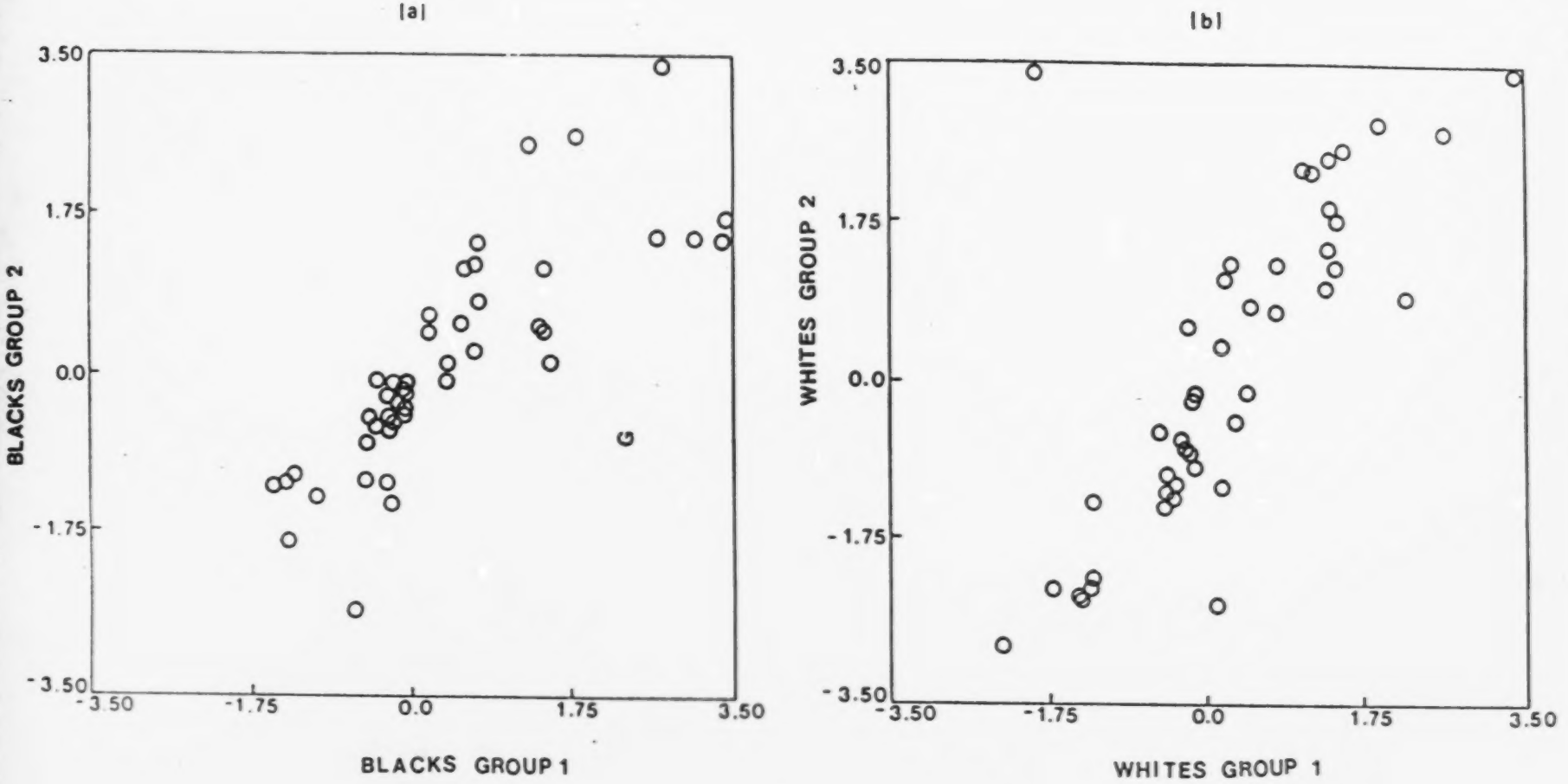

Figure 7. Stocking/Lord scaled Cleveland plots of item difficulty estimates in two equivalent black samples in (a) and white samples in (b).

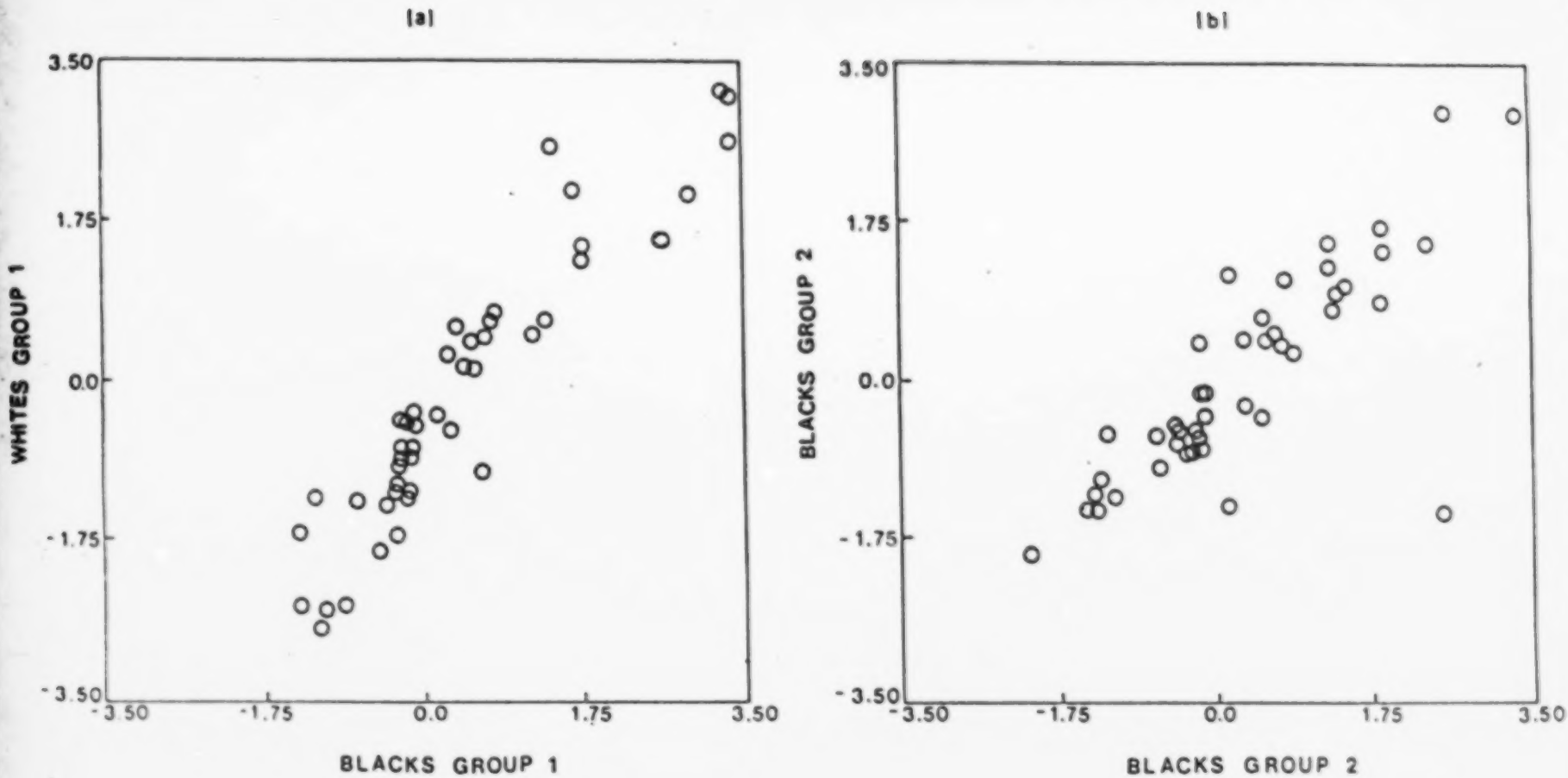

Figure 8. Stocking/Lord scaled Cleveland data plots of item difficulty estimates in the black and white samples (Sample 1 in a, Sample 2 in b).

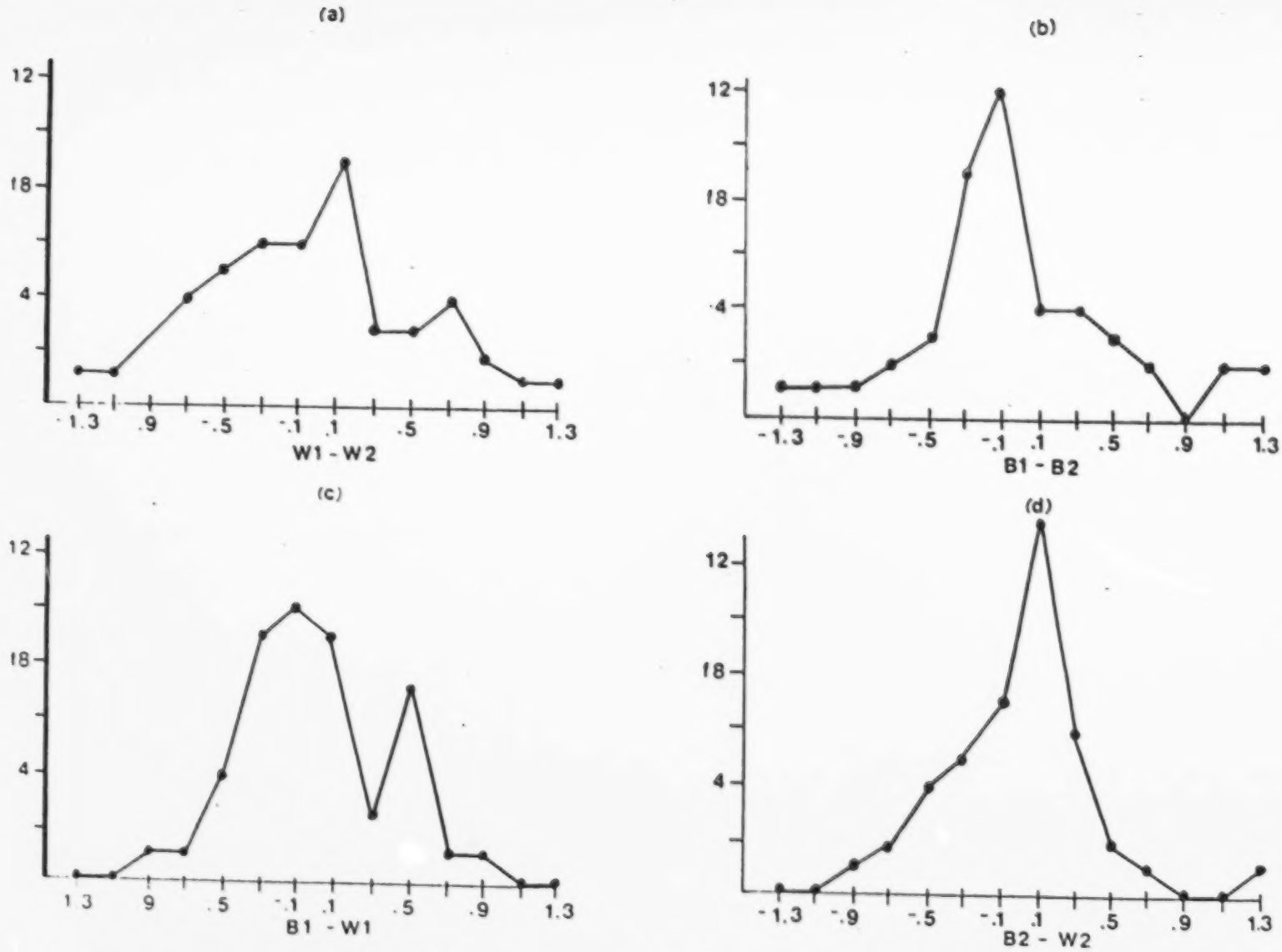

Figure 9. Distribution of b-value differences: (a) W1-W2, (b) B1-B2, (c) B1-W1, and (d) B2-W2.

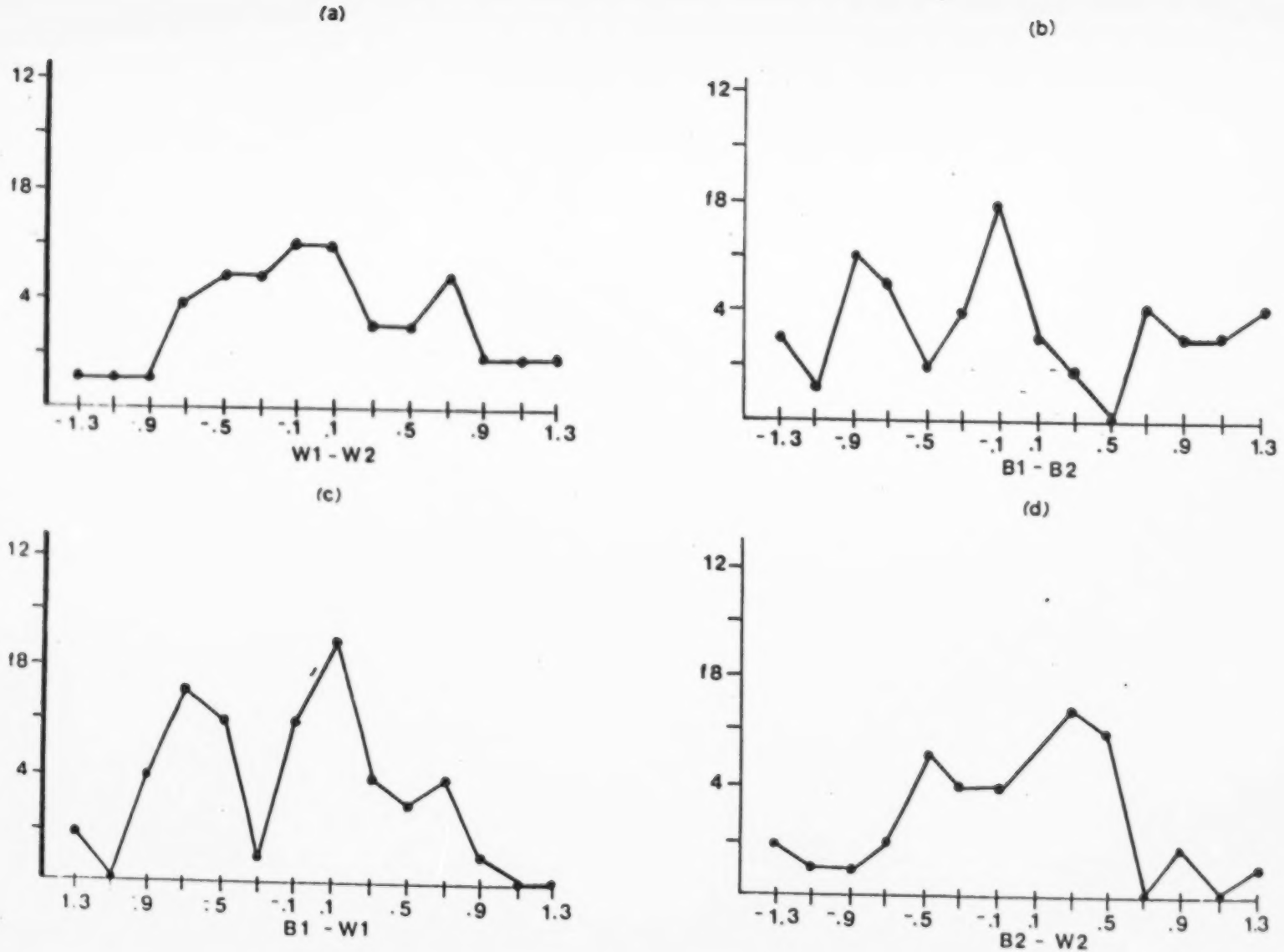

Figure 10. Distribution of standardized b-value differences: (a) W1-W2, (b) B1-B2, (c) B1-W1, (d) B2-W2.